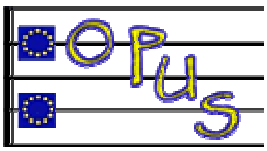


OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

WP6: Database Systems

Title : Generic Structures and Functionality for Support of Statistical Models in Statistical Databases –
Using Information from Statistical Models

Creator (Author):	Andrew Westlake	Survey & Statistical Computing
	Rajesh Krishnan	Centre for Transport Studies
Contributor :	Saikumar Chalisani	ETH
	Mike Collop	Transport *for* London
	Miles Logie	Minnerva

Identifier :	Deliverable D6.1 of project IST-2001-32471
Status :	Draft version of programme deliverable
Type :	Report
Version :	4.1 – Final Revision
Date.Created :	29 April 2005
Date.Modified :	12 July 2005
Date.Next version due :	
Submission Date :	
Subject.Category	
Subject.Keyword	
Source	
Relation	This header section draws on the e-GMS structure for document metadata developed by the UK e-Gov initiative.

Contract Date :	April 2003
Publisher (Project Coordinator) :	Imperial College London

Contact Person :	John Polak
Address :	Centre for Transport Studies Department of Civil and Environmental Engineering Imperial College London (South Kensington campus) London SW7 2AZ United Kingdom
Telephone :	+44-(0)20-7594.6089
Fax :	+44-(0)20-7594.6102
e-mail :	j.polak@imperial.ac.uk
Consortium :	CTS, TFL, KATALYSIS, ETHZ, FUNDP, PTV, SYSTEMATICA, WHO. MINNERVA, SURVEY & STATISTICAL COMPUTING, OXFORD SYSTEMATICS

TABLE OF CONTENTS

Technical Abstract 5

Executive Summary 8

1. Introduction and Framework 9

1.1 About OPUS 9

1.1.1 Background 9

1.1.2 Objectives of the OPUS project 10

1.1.3 Motivation 11

1.1.4 Subject areas 12

1.2 OPUS Project Work Package WP6 12

1.2.1 Objectives 12

1.2.2 Description of work 12

1.2.3 Deliverables 13

1.2.4 In Practice 13

1.3 Objectives of Deliverable D6.1 14

1.4 Structure of the Deliverable 14

2. Background 15

2.1 The LATS Database environment 15

2.2 Functionality in Nesstar 15

2.3 Statistical Database Enhancement 16

2.4 Supporting Models in the London and Zürich Test Cases 16

3. Results from a Statistical Model 18

3.1 Outputs from the Opus Methodology 18

3.2 Why is Synthetic Data Different? 19

3.3 Provenance and Reliability 20

3.3.1 Objectives 20

3.3.2 Model Form 21

3.3.3 Data used 21

3.3.4 Parameters 21

3.3.5 Domain displays 21

4. Issues in this report 22

4.1 Issues 22

4.2 Areas of Development 22

4.2.1 Model Forms 22

4.2.2 Model Outputs 22

4.2.3 Metadata Structures 23

4.2.4 Metadata Functionality 23

4.2.5 Information Requirements 23

4.2.6 Display Components 24

4.2.7 Presentation Templates 24

4.2.8 Implementation Architecture 24

5. Development Areas 25

5.1 Information Requirements 25

5.1.1 Basic Areas 25

5.1.2 Domain Users 25

5.1.3 Creating Awareness 25

5.1.4 Presentation 26

5.2 Representation of Models and Methodology 27

5.2.1 Static and Dynamic elements 27

5.2.2 Types of Statistical Model 27

5.2.3 Domain Models 27

5.2.4 Fitting Methodology 28

5.3 Evaluation of Model Reliability 28

5.3.1 Bayesian Model Checking 28

5.3.2 Distributional Displays 28

5.3.3 Parameter Reliability 29

5.4 Display Components 29

6. Architecture 32

6.1 General Approach 32

6.2 Technology components 32

6.2.1 Nesstar 32

6.2.2 R and Rserve/RCgi 32

6.2.3 HTML and XHTML powered by Apache 33

6.2.4 VISUM 33

6.3 High-level architecture 33

6.3.1 Web based system 33

6.3.2 Integrated VISUM-R environment 35

References 36

Index 37

TECHNICAL ABSTRACT

Using Information from Statistical Models

This deliverable D6.1 is a result of Work Package WP06 of the OPUS project. Work Package WP06 has as title: "Database Systems". Its original objective was to explore ways to extend statistical databases to support statistical modelling, with associated metadata. The ideas proposed were to be implemented in the context of the 'LATS Database'

However, the LATS database does not exist in the form envisaged when the project proposal was written. Instead, we have Romulus, which is based on Nesstar, which is a Federated Database system for the dissemination and analysis of statistical data and results. This uses DDI for metadata about datasets, with some potential extensions. Nesstar is also used by ETH for Zürich data, where they have proposed various extensions for handling transport-related data structures.

There is no proposal to extend either of the Nesstar implementations to include Opus-style modelling within the dissemination systems, as was originally envisaged. However, there is the intention to use the Opus methodology outside the system, and to then include synthetic (simulated or enhanced) datasets. So, for the metadata and database contributions from WP06 we propose to concentrate on the support of the use of the synthetic data and other information obtained from the statistical model.

The term we use for this extension to the existing functionality is
Provenance and Reliability.

The Opus methodology is Bayesian, so all the information about a statistical model lies in the model specification plus the posterior distributions of the parameters. That is, all information about the underlying real-world system that is contained in observed datasets (and is pertinent to the model formulation) has already been extracted by the model fitting process, into the posterior distributions. In theory it is then sufficient to present just this extracted information (the model formulation together with the posterior distributions) to users. In practice, this will be too complex or impenetrable for most users, so, as with most statistical analyses, other forms of interpretation and presentation will be needed.

We anticipate the presentation of three forms of information derived from a model.

1. **Conclusions**. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.
2. **Estimates**. Presentation of the posterior distributions of quantities of interest from the underlying system. This can be done by showing summary statistics (particularly means and standard deviations) of the posterior distributions, or by presenting complete distributions, displayed as histograms or multivariate contour plots (for example). Note that the distribution represents our uncertainty about the true value of the quantity, so it is important to present this as well as any point (best) estimates.
   Population parameters of direct interest to users (for example, in decision mak-

ing) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process. But estimates can be obtained for any derivable measure on the underlying system, with a corresponding derived posterior distribution.

3. **Synthetic data**. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.
   The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations on the underlying system, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences.

These three types of information have close parallels with information obtained by more traditional methods. The difference is in the central role of the model in our methodology. Instead of presenting information that is directly derived from a dataset, and which is then inferred to be directly about the underlying system, all our information is mediated by the model.

The central role of the model is valuable because it allow us to generalise from actual data to all situations covered by the model. All information that we present will be valid information about the model, but will only provide useful insights about the underlying system if the model has a valid (and sensible) structure and is well-determined by the available data. Thus a user of information from the model should reasonably ask about the form of the model, the processes by which it was fitted, and the extent to which conclusions are well-determined.

We thus propose that two additional types of information should be available with all results that are derived from a statistical model.

4. **Provenance**. Information about the structure and objectives of the model (including its mathematical form), and about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.

5. **Reliability**. This corresponds to the Estimates topic above in that it relates to the posterior distributions of the model parameters. But instead of focussing on estimates of quantities of interest in the underlying system, it focuses on the uncertainty that remains about the model parameters.
   It is important to distinguish between *uncertainty* about parameters (which should generally decrease as more data is used or as the model formulation is improved) and *variability* in observed data that is associated with measurement processes or variability of behaviour.
   In particular, we are interested in how uncertainty evolves from the prior as-

sumptions through the various data fitting steps, and in how this uncertainty feeds through into uncertainty about measures on the underlying system.

The source of most of this information is the metadata that describes a statistical model and that records (like an audit trail) the processes used to arrive at the posterior distributions in the final state of the model. High-level structures for this metadata have been described in deliverable D3.1, and we expand on these later. We also address issues about the delivery and presentation of this information to users.

EXECUTIVE SUMMARY

Using Information from Statistical Models

This document is Deliverable D6.1 of the Fifth-Framework project (FP5) OPUS. The OPUS project aims to develop and demonstrate statistically sound methods of combining datasets, where each provides partial information on a single complex of underlying variables.

The expected practical result of application of the OPUS methodology is a calibrated probabilistic model of the problem domain at hand, with which it is possible to calculate the most likely values of missing, unobserved, or unobservable quantities of the object system under study, with potentially important savings of time and resources.

This report discusses the requirements for using the results of the Opus Methodology in the test cases in London and Zurich. It has different objectives from those originally conceived, because the nature of the systems that will be used for the test cases has changed, particularly in London.

After discussing the current context for the test cases, the report addresses the issue of providing information to support the use of the results from applying the Opus Methodology to the test cases. The task is to provide the user with information to inform and support their use of results obtained from a calibrated model. This focuses on information about the ***provenance*** of the model, that is, how it was constructed and how it was calibrated, and the ***reliability*** of the estimates obtained from it, which relates to the posterior distributions of the parameters of the model.

The tasks to be addressed are discussed, and a broad vision for the nature and architecture for the solution is presented.

1. INTRODUCTION AND FRAMEWORK

1.1 About OPUS

1.1.1 Background

OPUS is a large information management research project, supported by Eurostat as part of the European Commission's Information Society Technologies (IST) Programme. The overall aim of the OPUS project is to enable the coherent combination and use of data from disparate, cross-sectoral sources, and so contribute to improved decision making in the public and private sector within Europe. The research is focused on developing an innovative methodology, incorporating statistical and database systems. Transport planning is a prominent example of a topic that uses multiple sources of data, and will be the main test case for OPUS, but the cross-sectoral nature of the research will be demonstrated through the inclusion of an application in the field of health information as another example.

To meet the needs for comprehensive information on socio-economic systems such as urban and regional transport planning, and in the health services sector, data from diverse sources (e.g. conventional sample surveys, census records, operational data streams and data generated by IST systems themselves) must be combined. There is currently no appropriate developed methodology that enables the combination of complex spatial, temporal and real time data in a statistically coherent fashion. The aim of the project is to develop, apply and evaluate such a methodology. OPUS will develop a general statistical framework for combining diverse data sources and specialise this framework to estimate indicators of mobility such as travel patterns over space and time for different groups of people. The project will undertake pilot and feasibility study applications in London, Zurich, Milan, and on a national level in Belgium. Methods for extending the framework to information aspects of the health domain will also be investigated.

The benefits of OPUS will be:

- Improved estimation of detailed travel demand, using all available information;
- Avoidance of simplified combination of data that can give erroneous estimates;
- Indicators of data quality, to provide guidance for new data collection;
- A framework for managing data from rolling survey programmes;
- Better understanding of the role of variability and uncertainty in results and models;
- Avoidance of confusion from different, apparently conflicting, estimates of the same quantity;
- A generalised methodology for other domains of interest.

The participants in the OPUS project are as follows:

Research Organisations

- CTS (Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London), United Kingdom – Lead Partner
- DEPH (Department of Epidemiology and Public Health, Imperial College London), United Kingdom
- ETHZ (Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau), Switzerland
- FUNDP, Transport Research Group (Facultés Universitaires Notre-Dame de la Paix), Belgium

Practitioners

- Minnerva Ltd., United Kingdom.
- Survey and Statistical Computing, United Kingdom.
- Katalysis Ltd., United Kingdom.
- PTV AG, Germany
- Systematica, Italy.
- Oxford Systematics, Australia: Peer Reviewer

Public Bodies

- Transport for London (TfL), United Kingdom.
- World Health Organisation (WHO), Italy.

1.1.2 Objectives of the OPUS project

To meet the needs for comprehensive information on socio-economic systems such as urban and regional transport planning, and in the health services sector, data from diverse sources (e.g. conventional sample surveys, census records, operational data streams and data generated by IST systems themselves) must be *combined*. There is currently no appropriate developed methodology that enables the combination of complex spatial, temporal and real time data in a statistically coherent fashion.

The overall aim of the proposed project is to develop, apply and evaluate such methodologies, taking as a specific case study the transport planning sector. The specific objectives of the study are:

- To develop a generic statistical framework to enable the optimal combination of complex spatial and temporal data from survey and non-survey sources. This framework will specify how to optimally estimate the underlying population parameters of interest taking into account the structural relationships between the different measured data quantities and the sampling and non-sampling errors associated with the respective data collection processes. It is envisaged that the framework will be broadly Bayesian in nature. The framework will make no specific assumptions regarding the particular structural and sampling/non-sampling errors and will thus be relevant to a wide range of application domains.
- To apply the generic framework within the field of urban and regional transport planning. This will involve the definition of specific structural relationships

amongst measured quantities and the characterisation of sampling/non-sampling errors, based on domain knowledge from the field of transport planning.
- To develop the necessary database and estimation software to enable the application of the statistical framework in a number of case study areas.
- To undertake a major pilot application study in London, focusing on the derivation of indicators of the mobility and the performance of transport policy measures.
- In parallel, to investigate the feasibility of applying the framework and methodologies developed both in other transport planning contexts and in other proximate domains, specifically environmental management and social statistics.
- Based on the experience gained in the pilot application and the feasibility studies, to evaluate the performance of the proposed methods and to define the scope and approach for wider applications in relevant domains including environmental management and health care.
- To disseminate the results to the relevant academic and practitioner communities.

1.1.3 Motivation

OPUS addresses the situation in which the analyst must combine data from a variety of different data sources to obtain a best estimate, or a fuller understanding, of a system. Such a situation can arise for a number of reasons including:

- No single source contains sufficient information by itself; or
- Multiple sources naturally arise (e.g. through observations at different levels of spatial or temporal aggregation or by means of different survey methods), resulting in a need to reconcile potentially conflicting estimations; or
- The need to update or transfer an existing set of data and parameter estimates when additional information becomes available.

Problems of combining data from different sources to produce consistent estimates of underlying population parameters arise in many fields of study including environmental monitoring, epidemiology and public health, earth observation, geographic information and navigation systems, transport and logistics, and economic and social statistics. Although the risks of using *ad hoc* combination rules and procedures are well understood, there are nevertheless many examples from practice in which just such approaches are still used. This reflects the fact that, although relatively straightforward methods exist for simple cases, there does not exist a coherent and well developed set of applicable methods capable of dealing with the full range of data combination problems, including factors such as:

- Data sources that provide both direct and indirect information on the relevant population parameters
- Data that are presented at different levels of aggregation
- Data sources with differing levels of statistical precision or user confidence
- Data that overlap, but that may provide different or conflicting information
- Gaps in the data observations
- The issues raised by the aging of sample survey data and the consequent need for updating
- Accommodating the updating sources
- The effect of sampling and non-sampling errors (including survey non-response and other sources of missing data)

- The opportunities presented by new data streams from IST systems

The key scientific objective of the project is to develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey and non-survey sources. The framework will be sufficiently abstract to be applicable to a wide range of potential domains.

Associated with this overall objective is the need for a suitable representation of the statistical metadata that is used for the specification and application of such a framework. That is the immediate objective of this report.

1.1.4 Subject areas

OPUS provides a generic approach but, in each case, it is necessary to make this approach specific to the particular area of interest (whether the area is geographical or topical in nature). A particular test-bed is transport in London, but studies will be made for transport in Belgium, Switzerland, and Italy, as well as health studies.

1.2 OPUS Project Work Package WP6

This section summarises the specifications for this work package, as taken from the project proposal.

1.2.1 Objectives

- *Building on the work done in WP3 and on the previous design and implementation work for the LATS 2001 project in London, to establish designs and implementation criteria for supporting modelling within statistical databases.*
- *To undertake the implementation of the database and metadata software in support of WP7 and the applications in WP8 and WP9.*

Note that this implementation has to be done in the context of the pilot applications in London and Zurich, (within WP8 and WP9) but we will also investigate whether it is possible to do an implementation in a standard generic environment. The former will be sufficient for supporting work packages 6 and 7, but the latter would be more valuable in that it would be more easily transferred (and extended) for other domains. The database environment plus the modelling software would provide a more general-purpose modelling environment. However, we are well aware that producing general-purpose facilities is very difficult, so, given the time and resources allocated to this work package, we will only attempt this generalisation if a particularly straightforward solution presents itself.

1.2.2 Description of work

The work programme will consist of two inter-related activities.

First, extensions to the existing object model developed for the LATS 2001 project in London to be more generic and to provide support for and an interface to a separate modelling component. This will include enhancement of the LTS model to provide the metadata extensions developed in WP3, and design of suitable functionality (including interfaces) to support modelling. This will be done in a more generic way, so that the model is applicable more widely than just in the LATS 2001 (or even transport in general) context

Second, the design and implementation of a test database environment to support the modelling software to be developed by WP7. This implementation has to be done in the LATS 2001 project for London, but this will also include study of whether it is possible to do the implementation in a standard generic environment.

1.2.3 Deliverables

D6.1 Report on the Database Structures and Functionality For Generic Support
D6.2 Report on the Implementation of Modelling Support in Statistical Databases
D6.3 Database System Enhancement

Because any information that is fed back into the statistical system will be a result of the Opus methodology, we have decided to focus our efforts in WP6 on supporting the use and understanding of this information. This will be done alongside the existing statistical system, using the same technology where appropriate, and implemented in such a way as to appear as seamless as possible for the user of the statistical system. A specific objective of this enhancement will be to implement new facilities in a way that demonstrates their usefulness and facilitates any later implementation within the Nesstar system.

1.2.4 In Practice

The proposal for the Opus project was written shortly after the completion of a design report for a Statistical Database for the results of the 2001 London Area Transport Survey (see [West01]). This report envisaged the construction of a specialised database system.

After a more detailed investigation T*f*L decided that the full implementation of a new system was neither economical nor practical, and instead have constructed a system (called *Romulus*) which is based on the dissemination package Nesstar[1]. This is the system that will be used to host suitable results produced by WP8 (the London Test Case), and so is the target system for WP6. WP9, the Zürich Test Case is also to be used with a system constructed using Nesstar, so the results of WP6 will be usable with both test cases.

This situation has been discussed already in D3.2. We concluded there that the focus of WP06 had to be been changed, because the inclusion of modelling methodology and metadata in statistical databases is no longer envisaged. Instead, we have decided to concentrate on issues related to making the results of modelling available in a statistical database context. This includes discussion of the way in which metadata about the model fitting process can be used to inform users' of the model results. We focus on the broad requirements for making use of the results of the Opus methodology (or other statistical methodology) in association with statistical databases, and on the overall architecture needed to support those requirements.

We refer to this problem area as *Provenance and Reliability* with the implications of this term being expanded in later sections.

---

[1] The Nesstar product was developed under a number of EU framework projects, including the Nesstar and Faster projects. Commercial exploitation is being undertaken by Nesstar Ltd, a spin-off company hosted at the University of Essex – see www.nesstar.com.

1.3 Objectives of Deliverable D6.1

The key scientific objective of the Opus project is to develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey and non-survey sources. The framework will be sufficiently abstract to be applicable to a wide range of potential domains.

Associated with this objective is the need for a suitable representation of the statistical metadata that is used for the specification and application of such a framework. That is the objective of deliverable D3.1. In D3.2 we discussed the needs and requirements of the LATS group for the enhancement of their statistical dissemination system to make use of the Opus methodology, and the implications of those requirements in terms of functionality and implementation strategy.

The current report elaborates the work of WP03. We discuss in detail the requirements for supporting users of model results, at a generic level. We also set out how we propose to implement such functionality (in conjunction with WP07) for use in WP08 and WP09.

1.4 Structure of the Deliverable

Chapter 2 presents background information, describing the current situation with respect to the existing statistical dissemination systems in use in London (at T*f*L for the LATS data) and in Zürich and setting out the basic objectives for WP6. Chapter 3 describes the requirements, focussing on the support of the use of synthetic data. In Chapter 4 we identify areas that need to be addressed to work towards our objectives, and Chapter 5 elaborates on some of these. Chapter 6 discusses a possible architecture for the implementation of the proposed functionality.

2. BACKGROUND

2.1 The LATS Database environment

The proposal for the Opus project was written shortly after the completion of a design report for a Statistical Database for the results of the 2001 London Area Transport Survey (see [West01]). This report envisaged the construction of a specialised database system:

> *The LATS database system is intended to be a dynamic resource containing information about travel in London. It will contain information about demand, use and attitudes and will cover all modes of transport. It is complementary to various other databases about transport facilities in London, and will have facilities to co-operate with them.*
>
> *This document presents various issues relating to the design and use of such a database. It considers the objectives of the database, the use and users of the database, the main requirements for features and functionality (with considerable detail in some areas), and some technologies relevant to the implementation. It is the main output from a design study to investigate the architectural and functional characteristics required for the database.*

After a more detailed investigation T*f*L decided that the full implementation of a new system was neither economical nor practical, and instead have constructed a system (called *Romulus*) which is based on the dissemination package Nesstar. This is the system that will be used to host suitable results produced by WP8 (the London Test Case), and so is the target system for WP6. WP9, the Zürich Test Case is also to be used with a system constructed using Nesstar, so the results of WP6 will be usable with both test cases.

2.2 Functionality in Nesstar

The Nesstar system manages user access to a distributed database of statistical data. It also provides basic statistical analysis facilities.

Statistical data in Nesstar is stored either as normal data records (referred to as Survey Data), or aggregated into multidimensional data cubes (called Summary Data). All datasets are described using the DDI Codebook metadata standard (see [DDI]), which covers both datasets and the variables within them. The metadata is stored in XML documents.

The Nesstar system uses an enhanced web-browsing interface to present information to the user. This uses specialised components to display specialised information, but provides the flexibility to display any information that can be formatted as or linked into an HTML document. The linking is important, as it is designed to provide access to information that is not stored within the Nesstar system, but can be viewed from within the Nesstar interface.

This functionality provides the basic building blocks for the extensions to be produced under WP6.

2.3 Statistical Database Enhancement

Nesstar is not a partner in the Opus project. While the project has some aspirations that aspects of the Opus methodology might be integrated with Nesstar (and other, similar systems), this would be a matter for an exploitation phase subsequent to the completion of the Opus project.

We thus do not expect to integrate any parts of the Opus methodology directly into the existing statistical dissemination databases during the lifetime of the project (though it may happen later). In particular, we do not expect to implement any of the Opus model fitting methodology within the systems. Instead, the model fitting will be done externally. Datasets (and the related metadata) will be used from the statistical databases (possibly later by direct extraction, but certainly initially by taking copies), and new information obtained from the modelling will feed back into the database.

Because any information that is fed back into the statistical system will be a result of the Opus methodology, we have decided to focus our efforts in WP6 on supporting the use and understanding of this information. This will be done alongside the existing statistical system, using the same technology where appropriate, and implemented in such a way as to appear as seamless as possible for the user of the statistical system. A specific objective of this enhancement will be to implement new facilities in a way that demonstrates their usefulness and facilitates any later implementation within the Nesstar system.

In consequence, this report is not (as originally envisaged) a specification for the inclusion of modelling methodology and metadata in statistical databases. Instead it addresses issues related to making the results of modelling available in a statistical database context. This includes discussion of the forms of information about the real-world system that will be used following the fitting of a statistical model, and of the way in which metadata about the model fitting process can be used to inform users of the model results. We focus on the broad requirements for making use of the results of the Opus methodology (or other statistical methodology) in association with statistical databases, and on the overall architecture needed to support those requirements.

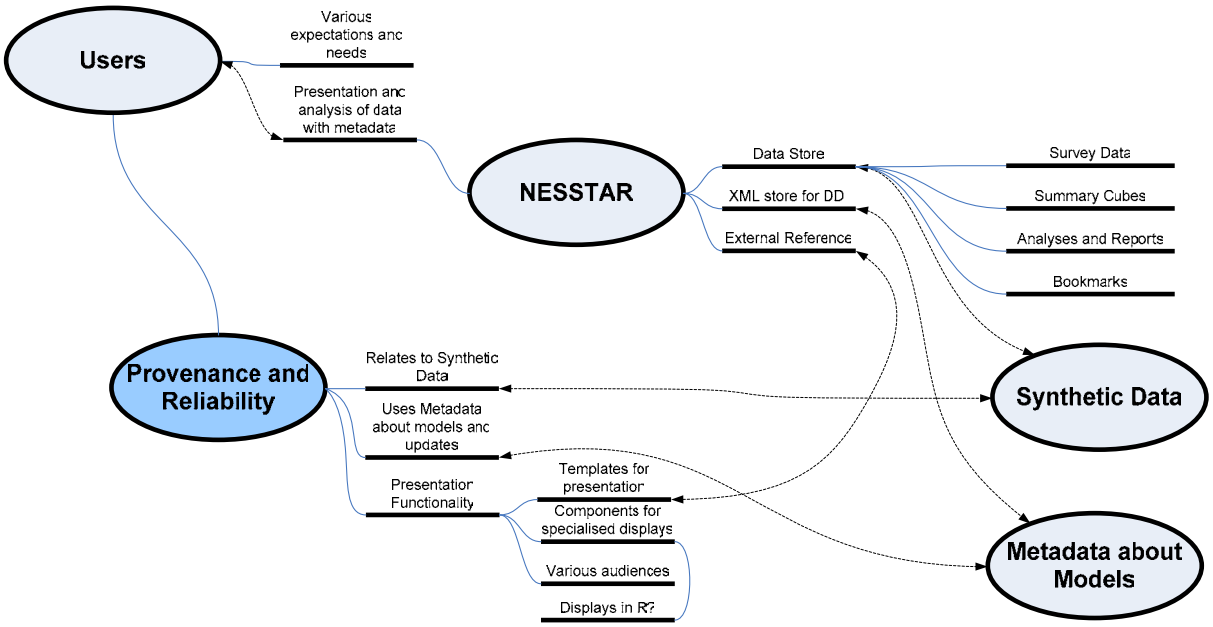
We refer to this problem area as *Provenance and Reliability* with the implications of this term being expanded in later sections.

2.4 Supporting Models in the London and Zürich Test Cases

There is no proposal to extend either of the Nesstar implementations (in London and Zürich) to include Opus-style modelling within the dissemination systems, as was originally envisaged. However, there is the intention to use the Opus methodology outside the system, and to then transfer information back into the system, in the form of both interpreted conclusions and synthetic (simulated or enhanced) datasets. So, for the metadata and database inputs we propose to concentrate on the support of the use of the synthetic data. Some of this information can be input directly into the existing systems, and some will (at least initially) be accessible from a separate web-based system operating alongside the existing Nesstar services.

Because the synthetic data looks like real data, no special facilities are needed to add it to the existing systems, or to use it within them. However, synthetic data is different, and to use it effectively (and correctly) the user needs access to additional information (metadata) about the process by which the data was synthesised from the model, and about the form and quality of the model. The structure for metadata about statistical models already proposed in deliverable D3.1 includes (potentially) all this information (it is in effect a complete audit trail for all the specifications and stages used to produce the synthetic data), but not in a form that will be readily accessible to users. So we will need to find ways of presenting this additional information that are accessible and comprehensible for different groups of user.

The following diagram shows some of the elements involved, and how they link into existing functionality within the Nesstar system.



While it is clear that the synthetic datasets can be stored within the Nesstar data store, the diagram is not intended to imply that other components will (necessarily) be closely integrated with Nesstar. It is more a question of exploiting synergies.

Nesstar uses XML for the storage of the DDI metadata, and we will (almost certainly) also use XML for the model metadata. Thus we will be using the same technology to access and manipulate the metadata, but will probably not make any attempt at this stage to integrate the model metadata with the DDI in the Nesstar metadata store. However, given the characteristics of XML and DDI, this should not be too difficult to do at some later stage.

Similarly, the presentation functionality (to provide access to information from the metadata about the synthetic datasets) will be implemented using web facilities, but will be hosted separately and linked using the external reference mechanism in Nesstar.

3. RESULTS FROM A STATISTICAL MODEL

3.1 Outputs from the Opus Methodology

The Opus methodology is Bayesian, so all the information about a model lies in the model specification plus the posterior distributions of the parameters. That is, all information about the underlying real-world system that is contained in observed datasets and is pertinent to the model formulation has already been extracted by the model fitting process into the posterior distributions. In theory it is then sufficient to present just this extracted information (the model formulation together with the posterior distributions) to users. In practice, this will be too complex or impenetrable for most users, so, as with most statistical analyses, other forms of interpretation and presentation will be needed.

We anticipate the presentation of three forms of information derived from a model.

1. **Conclusions**. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.
2. **Estimates**. Presentation of the posterior distributions of quantities of interest from the underlying system. This can be done in terms of summary statistics (particularly means and standard deviations) of the posterior distributions, or of complete distributions, presented as histograms or multivariate contour plots (for example). Note that the distribution represents our uncertainty about the true value of the quantity, so it is important to present this as well as any point (best) estimates. Population parameters of direct interest to users (for example, in decision making) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process. But estimates can be obtained for any derivable measure on the underlying system, with a corresponding derived posterior distribution.
3. **Synthetic data**. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.
   The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations on the underlying system, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences.

These three types of information have close parallels with information obtained by more traditional methods. The difference is in the central role of the model in our methodology. Instead of presenting information that is directly derived from a dataset, and which is then inferred to be directly about the underlying system, all our information is mediated by the model.

The central role of the model is valuable because it allow us to generalise from actual data to all situations covered by the model. All information that we present will be valid information about the model, but will only provide useful insights about the underlying system if the model has a valid (and sensible) structure and is well-determined by the available data. Thus a user of information from the model should reasonably ask about the form of the model, the processes by which it was fitted, and the extent to which conclusions are well-determined.

We thus propose that two additional types of information should be available with all results that are derived from a statistical model.

4. **Provenance**. Information about the structure and objectives of the model (including its mathematical form), and about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.
5. **Reliability**. This corresponds to the Estimates topic above in that it relates to the posterior distributions of the model parameters. But instead of focussing on estimates of quantities of interest in the underlying system, it focuses on the uncertainty that remains about the model parameters. In particular, we will be interested in how the uncertainty evolves from the prior assumptions through the various data fitting steps, and in how this uncertainty feeds through into uncertainty about measures on the underlying system. It will be important to distinguish between *uncertainty* about parameters (which should generally decrease as more data is used or as the model formulation is improved) and *variability* in observed data that is associated with measurement processes or variable behaviour.

The source of most of this information is the metadata that describes a statistical model and that records (like an audit trail) the processes used to arrive at the posterior distributions in the final state of the model. High-level structures for this metadata have been described in deliverable D3.1, and we expand on these later. We also address issues about the delivery and presentation of this information to users.

3.2 Why is Synthetic Data Different?

A statistical model consists of a mathematical specification of relationships between variables, probability distributions that capture the variability in these variables, parameters that describe the distributions and (possibly) the relationships, and probability distributions that capture the uncertainty in our knowledge about the parameters. By sampling from the probability distributions for uncertainty and variability, and by following through the mathematics of the relationships, we can generate an observation on the set of variables (and on the parameters). If we repeat this process we can generate a set of (synthetic) data records.

With real data collection, the sample size is of central importance (along with the survey design) in that it determines the amount of information we collect about the system being observed. The same is true of synthetic data, but the relevance is different, because the system being observed then is the model, not the underlying reality. The

sample size for synthetic data is arbitrary (or at least only limited by constraints of time and storage). A larger sample gives us better estimates about the parameters of the model, but that is not what is really of interest. What matters is the information **in the model** about the underlying reality, and that stays the same, whatever the size of the synthetic sample. We will use different criteria to determine the sample size to use. For example, we might choose to synthesise a complete set of all trips made in a particular time interval. We may choose to synthesise a dataset that corresponds to a real survey dataset. Or we may choose to synthesis multiple sets of observations so that we can explore variability between datasets (corresponding to parameter uncertainty) as well as within.

An estimate of a mean from the synthetic data will (probably) be an unbiased estimate for the mean of the posterior distribution of the parameter corresponding to this mean in the model. But the precision (standard error) of the synthetic mean will not tell us much about the precision with which the parameter is determined in the model. That information is in the model, and can be estimated from the synthetic data, but not with a naïve approach.

In general, analysis of relationships in synthetic data will provide some insight into the mathematical structure of the model, in a way that may well be more approachable than tackling the mathematics in the model directly. But to understand precision and uncertainty we need to draw users into other ways of looking at the information in the model.

In the transport domain, data is often processed through complex (transport) models to derive measures of direct interest, and the synthetic data may be subjected to such processing. It is difficult to conceive how to propagate information about model precision through such processes. However, if the measures derived are central to the domain, then they should be explicitly included in the model, and so information about them can be extracted directly, without the complex processing.

3.3 Provenance and Reliability

Here we summarise our objectives and approach.

3.3.1 Objectives

Users of synthetic data should properly be asking questions about how the data was generated and how much confidence they should have in conclusions drawn from it. We use the term *Provenance and Reliability* to refer to this area. This covers all issues to do with the understanding and interpretation of fitted models, not just those directly related to the use of synthetic data.

Different types of user will expect answers of different complexity and detail. Some answers can be generic, describing the philosophy behind the Opus methodology and Bayesian modelling, or showing (in UML diagrams?) the outline of the model fitting processes used. Other answers will need to be based on the specific components used in the model from which the data are synthesised, and further ones will make use of the detailed posterior information about the parameters. All this information will be available in the form of metadata, the top-level structure of which has been laid out in deliverable D3.1. The same information may need to be presented in dif-

ferent ways for different types of user. Not all reasonable questions will necessarily be amenable to being answered.

3.3.2 Model Form

For those interested in the specification of the models, we should be able to display various components at various levels of detail. This will extend from the top level GAPM applicable to the model, right down to the details of the mathematics involved in the relationships, constraints and distributions in a particular model. Some of this should be shown in mathematical form, but graphical representations will be used wherever possible. For models that fit the Graphical Models framework, a display similar to the 'Doodle' system in WinBugs will be used.

3.3.3 Data used

The model metadata includes links to all the data used in reaching the final calibration of the parameters, so this can be shown, and the user can explore the DDI metadata for datasets already within the Nesstar system. The links between variables in datasets and those in the model are also available.

3.3.4 Parameters

The final model state includes information about all the posterior distributions, for the (hyper) parameters, for those induced for the parameters of direct interest and for the variables. These can be presented using standard displays of distributions, probably using the facilities in R.

Such displays show the precision with which parameters have been determined. The reliability and suitability of the model can be explored through the progress of the parameter distributions through the calibration processes.

3.3.5 Domain displays

While generic displays of parameter distributions may be adequate for some statistical users, most practitioners are more used to working with specific forms of display that have been developed as particularly applicable to their domain of application. Here we face the challenge of enhancing such displays to show additional information about (particularly) reliability.

For example, in transport there are specialised displays, such as the network and OD diagrams produced by the Visum system. It is expected that these can be enhanced to show some aspects of variability and classification, and that they can be used to show appropriate parameter distributions, as well as distributions of actual traffic flows. This will be explored in association with WP7.

4. ISSUES IN THIS REPORT

In this chapter we build on material set out in the two deliverables from WP3, identifying at a generic level the structures and functionality that will be needed for the implementation within WP6.

4.1 Issues

Users of synthetic data and other model results (outputs) need access to additional information (metadata) about the process by which the results were generated (an audit trail) and about the quality of the models used, summarised in terms of the reliance that they can place in results obtained from the synthetic data. We thus face three issues.

1. Analysis of the information needed by different classes of user to support their use of model outputs.
2. Design of suitable metadata structures to capture and hold this information.
3. Design and implementation of suitable presentation material and functionality, so that users with different interests and skills can access and use the information that they need.

A start has been made on point 2 with the metadata structure for models proposed in D3.1. This receives further elaboration below, and we also address point 1. Point 3 is the subject of the remainder of WP6, and we indicate some general ideas for the architecture of this.

4.2 Areas of Development

In this section we identify and elaborate on specific issues that need to be addressed.

4.2.1 Model Forms

Many statistical models are generic, but some methodologies assume particular model forms. Common examples are the generic linear models, and the conditionally independent models used in Graphical Modelling. Further, particular application domains make assumptions about the form of statistical models used. Transport makes use of many things that it calls models but which are not statistical in nature. Are there constraints on the types of model we can handle, and if so, what are the restrictions?

4.2.2 Model Outputs

What information about a real-world system, based on a statistical model, do we expect to make available to users, and in what form?

We have already discussed this in the introduction and identified three types of information that are likely to be derived from a statistical model. These are Conclusions, Estimates and Synthetic Datasets.

4.2.3 Metadata Structures

What information is needed about the fitted statistical models and the processes and resources (including datasets) used for the fitting?

General structures have been identified in D3.1, but further elaboration and specificity is needed here.

A recent initiative related to R is an interface for Graphical Models (the gR package [HøDe04]). At an initial examination, this seems to adopt a similar abstract approach to the structuring of models as we have proposed in D3.1, though it is less generalised. An associated proposal for linked data and metadata covers the same basic ground as the triple-S standard[1], though in a more generalised way, including variable typing. These proposals are examined in more detail later.

Additional work is needed on the representation of the fitting process, particularly where the process involves procedures that form a 'black-box', and are not made explicit in the (statistical) model specification. This seems to be common in the transport domain, where 'transport modelling' is used for difficult tasks such as estimating link loading from Origin-Destination (O-D) flows.

Instances of the model outputs discussed previously (which will exist in their own right) will need links from (and to) the metadata structures for the model, so that information about the model and processing can be accessed along side the results. This need is already identified in D3.1, but needs elaboration.

4.2.4 Metadata Functionality

What manipulation of the metadata is needed?

This relates to the capture, transformation and presentation of information in the metadata store.

For example, the way in which a statistical model is specified in a particular statistical package is unlikely to be identical to the way we wish to represent that information. We assume that a mapping exists that does not lose any important component of the specification (if it did, our structure would be inadequate). So we need to consider how to capture the specification and how to be able to transfer it between the package and the store. In the early stages we will probably do this manually, but automatic transfer will be explored, at least at an abstract level.

Similarly, we will need to be able to access elements of the stored information and transfer it to the display components and information pages accessed by users. We assume that the data store actually used for instances of model metadata will be constructed using XML, so that transformations of the metadata to other forms can be achieved through the use of XSLT.

4.2.5 Information Requirements

What information can a user reasonably request, and what information should we display automatically whenever model-related information is referenced?

---

[1] See www.triple-s.org.

Under this heading we elaborate what we see as the various requirements for information to support the proper and correct use of the results from the statistical models. We also address concerns about how to bring the existence and importance of this information to the attention of the (possibly naïve) user of the model output.

4.2.6 Display Components

What are the generic display forms that are needed to present the required information?

On the assumption that information pages can be constructed by using a combination of textual information and graphical displays, we list and identify some of the generic components that we hope to use to contribute dynamically to these pages.

4.2.7 Presentation Templates

How are the various display components brought together to address particular needs for information?

Templates are intended to address particular needs at a general level, and will then be used to construct specific pages in particular contexts. Note that some such templates will be closely linked to the domain in which the methodology is applied.

4.2.8 Implementation Architecture

Probable methods for integrating all these elements have been identified, but will need to be worked out in detail. Nesstar should be able to link across to new facilities, and R should provide the host for the integration.

We discuss the general principles and structure, but (at this stage and for this report), try to remain above specific choices (except for illustration). These ideas will be developed in the later parts of WP6, where we do need to create an implementation.

5. DEVELOPMENT AREAS

This chapter contains elaboration of some of the problem areas identified previously. In some cases we have identified solutions, but often we merely elaborate our understanding of the problem.

5.1 Information Requirements

5.1.1 Basic Areas

We have already identified the three major areas of information about a model that need to be made available to users. These are:

1. The specification of the statistical model that has been fitted.
2. The audit trail of the processes and data used to fit the model.
3. The posterior distributions of the parameters of the model, which contain all the information about the model extracted from the data.

The presentation methods specific to these areas described elsewhere in this report will probably be sufficient for the user adept in statistical methods and mathematics. They can use these displays as tools and with them find answers to the questions that they themselves raise about the model.

5.1.2 Domain Users

For a user who is not familiar with statistical methodology (a non-specialist) we need to do more. We will need to provide displays that are simpler (and so do not rely on user understanding of abstract representations), and that are more focussed on the application domain of the user (so we may need different displays for different domains).

The more demanding problem, however, is that we cannot rely on the user being able to formulate appropriate questions, or even recognising that questions need to be asked. So we must address two problems.

1. How to create awareness in the user of the different nature of information obtained from a statistical model, and
2. How to provide a route map for the user through the potentially relevant questions.

For these users it is not enough to provide tools: we must provide solutions, from which they can assess the reliability of conclusions that they may want to draw from the information from the statistical model.

5.1.3 Creating Awareness

Within applications that we control and that provide information from statistical models, we are able to automatically introduce links and prompts to the additional

information about provenance and reliability. An example of this is discussed below in the context of transport networks in the Visum software.

Otherwise we rely on the original authors of the information to create awareness, rather than using software to do it automatically. Where the information from a model is used in an analysis, the analyst should already have made suitable investigations, and so can report these with the analysis and direct the reader to a suitable context for further investigation.

Where information is made available without commentary (and the major example of this is in synthetic datasets), it is necessary to make use of less direct methods, relying on the existence of metadata associated with the information. This will exist for synthetic datasets placed into a Nesstar system, where the DDI metadata allows extensive commentary to be associated with the dataset, but may not be possible in other contexts. In general, we rely on the creator of the information to take every opportunity to direct users to related information about provenance and reliability.

5.1.4 Presentation

Once we have the attention of the user we must guide them to understanding of the nature of statistical models in general, and the reliability of specific information in a particular context.

The planned approach is to develop a series of web pages that can act as a template for constructing a specific site that would support usage of a group of statistical models within some domain of application. The later parts of WP6 will elaborate these pages, and then contribute to implementing them in the context of the test cases in WP8 and 9.

Some generic requirements for these pages can be identified.

1. They must provide both general guidance and specific information about the model currently of interest. So some pages will be largely static and others will be based on the model metadata.
2. Many users will be interested only in part of the model, probably related to a particular output or component of the underlying system. It should be possible to quickly focus on the relevant parameters and data (the links are in the metadata) without loosing access to appropriate guidance.
3. Particularly in models with large numbers of structured parameters (as, for example, with origin-destination pairs in a transport system) we cannot expect the non-specialist user to explore the whole range. So it is desirable that the presentation system should be able to make some automatic assessment of the reliability of different parameters (or groups of parameters), or the indication of possible anomalies. Further exploration of the Bayesian literature is needed to try to identify suitable measures for this, though we are aware of the inherent danger is such search techniques.

5.2 Representation of Models and Methodology

5.2.1 Static and Dynamic elements

For any fitted model we will have metadata that specifies the detailed form of the model, plus the processes (and data) used to arrive at the posterior distributions that encapsulate the information in the model.

In representing the model, some elements will be based on the specific formulation and processes of the model, and so will need to be dynamically created for each different model. Other elements will relate to more standard procedures and methodologies, and these can be represented in a more static way, with references showing which procedures were used in a particular model.

5.2.2 Types of Statistical Model

For the moment we are basing our plans for WP8 and 9 on the standard Graphical Model (GM), with extensions to multiple data sources and feedback loops. This uses the basic MCMC fitting methodology (as implemented, for example, in WinBugs), and then iterates using an E-M (Expectation – Maximisation) approach.

The form of such models can be represented succinctly and formally using directed graphs. The E-M extension removes the restriction to acyclic graphs that is central to the GM methodology. This form of representation of models is currently used in the 'Doodle' component of WinBugs, and is relatively easy to implement using other graph drawing systems.

Such graphs are likely to be a useful form of presentation for the mathematically or statistically adept, but will not be sufficient for other users. Less formal diagrams (similar to graphs) will be useful here, but more wordy explanations will also be needed.

The approach here is intended to be generic, and so applicable to other forms of statistical modelling. However, it is unlikely that any other such forms will be explored within the scope of the Opus project.

5.2.3 Domain Models

Transport (in a way similar to various other domains) has a number of mathematical techniques that are used to estimate factors that are difficult to observe. These are referred to as models, but are different in nature from statistical models. They are usually deterministic (if complex) and can be thought of as allocation procedures, not estimation ones. An example could be a 'Route-Flow' model, which determines the sharing among different possible routes (with known cost functions) of the demand for travel between a set of origins and destinations.

Sometimes the processing involved in such models is widely understood and so can be treated as a 'black box', without further explanation. For other models the process will need to be represented using some form of process diagram. The tools available in UML should be sufficient for this, with the Activity Diagram (which corresponds to a flow-chart) being particularly useful.

5.2.4 Fitting Methodology

The procedures used to combine domain methods (models) and statistical methodology, using data to derive posterior distributions, also need to be represented. Since this is a process representation, UML is again seen as the main tool for this.

5.3 Evaluation of Model Reliability

5.3.1 Bayesian Model Checking

In all statistical modelling we face the problem of determining whether the chosen statistical model is well-suited to the reality that it represents, and whether it is well-determined by the fitting process that has been used. With classical methods we explore suitability by looking at residuals (comparing observed and fitted data values) and worrying about distributional forms (eg q-q plots), influence, etc. We address quality of fit by looking at measures such as the coefficient of determination (R^2), the residual variance and the significance levels of parameters.

Similar methods can be used in a Bayesian context, though not all have an immediate equivalent. In addition, however, we have the issue that the final posterior distributions are influenced by the initial information in the form and parameters of the prior distributions.

Gelman et al. [GCSR04] discuss model checking in a Bayesian context, and focus on the comparison of data distributions with the equivalent posterior distributions derived from the model. They point out in particular that it is not sufficient to have good correspondence in the mean and variance of a distribution – even with this it is possible to have a poor fit in the tails or inconsistent skewness – so they recommend examining whole distributions to assess model quality.

This is why we place heavy emphasis on the display of distributions as a means to understand model quality and parameter reliability

5.3.2 Distributional Displays

We take the statistical system R as our model for statistical functionality, and this can display the following distributional forms (not an exhaustive list).

1. Single distribution curves, derived either from a mathematical form or by smoothing an empirical (data) distribution.
2. Bi-variate distributions, as a (flat) contour plot or as a view of a 3-dimensional contour surface, where the view point can be dynamically moved.
3. Multidimensional point clouds, where the view point (for projection onto the 2-dimensional image) can be moved through all the dimensions.

These displays provide us with tools to investigate the posterior distributions of parameters, looking at values, shapes and inter-dependencies. They can also be used for any other distributions, such as the data distributions predicted by the model, or empirical distributions from datasets.

All of these displays can be used to compare distributions. For example, two (or more) distribution curves can be shown in the same diagram, or two (or more) data sources can contribute to a point cloud, using colour to distinguish the sources. Such displays form a major tool for model checking, because they allow us to compare observed distributions of data with simulated data from the model.

Comparison between displays can be done either manually (by producing more than one diagram) or automatically. The latter uses the Lattice Plot functionality of R, which produces the equivalent of a cross-tabulation in which each cell is a diagram. The Lattice dimensions (usually at least two) can be sub-groups associated with conditioning variables, such as geographical location (or zone) or demographic characteristic (e.g. occupation) or purpose of travel, or any other measure linked to the model.

5.3.3 Parameter Reliability

Where model fitting involves a single step, the posterior distribution contains all the information about the reliability of a parameter. However, where fitting involves a sequence of steps (for example, in an E-M iterative process), it is of interest trace the changes in the intermediate distributions. This may help to identify (for example) which steps or data sets contribute most to the determination of the distribution for a parameter.

We do not yet have a specific proposal for the best way to extend the distribution displays already described to the display of sequences of parameter distributions. Further work will be undertaken to address this requirement.

5.4 Display Components

On the assumption that information pages can be constructed by using a combination of textual information and graphical displays, we list and identify some of the generic components that we hope to use to contribute dynamically to these pages.

The following display types are needed to represent the information about provenance and reliability:

1. Spatial context of information: The key aspects of the information generated by the Opus methodology should be displayed in an appropriate spatial context. This can be done by enhancing a domain specific application such as Visum (which already has functionality to display network and spatial aspects of data) to handle reliability and other distributional information from the model.

   As an example, let us say that we want to display the estimated traffic flow over a network and the associated reliability of the estimation represented as a probability distribution. The expected value of the flow can be represented by the width of the link in a Visum display. While the expected value can be communicated easily, it is important that reliability of the estimate is made obvious graphically even to a naive user. This can be done in the following ways.

   a. **Colour grading**: In this scheme, colour scale across the width represents the probability of the expected value. For example, for an estimate that is more reliable, the colour could be green around the width equivalent to the ex-

pected value and red for values that are less likely. For a less reliable estimate, the width of the link will be larger with lesser greenish hue (than used for a more reliable estimate) around more probable values. Alternatively we may be able to use colour intensity to indicate lower confidence in values. This colour mechanism should be supplemented by a click-activated pop-up displaying the statistical distribution of the estimated value.

b. **Pane**: Whenever modelled data is displayed by Visum, a pane should appear in a corner of the application, displaying a summary of the statistical properties of the modelled data. Users will not be able to turn off this pane while modelled data is displayed.

c. **Mouse-over**: When a user hovers the mouse pointer over a network link, a transient window will display the statistical properties of the estimated distribution graphically. Though this provides a cleaner interface, the reliability aspect of the synthetic data is not made obvious to the user unless the user places the mouse pointer over the links.

d. **Click**: In this scheme, the user does not see the reliability aspect of the statistical data unless he/she explicitly requests the information through mouse clicks. However, the user's attention can be drawn to the reliability aspect by annotating the links with the value of the variance (similar to marking the lengths in structural drawings). A conspicuous annotation will draw the user's attention and invite a click-through for viewing further information about the data.

2. Audit trail of provenance: Audit trail describes how a synthetic data set is produced. The meta-data about a synthetic population will link to the underlying data that is used to create the population, as well as the mathematical model, parameters and the process followed in combining them. It is assumed that this information is disseminated through a web based interface. This information is best represented in a the following ways:

e. **UML diagrams**: State-chart diagrams in UML provide a useful way to represent processes, states and state transitions, and are well suited to representing the provenance of synthetic data graphically. The state-chart UML diagram should give a user an end-to-end picture of how synthetic data is produced, including the models and processes involved. The user will be able to zoom in on any aspects of the state-chart diagram to gain a deeper understanding of a particular dataset or a mathematical model.
The basic elements of this state-chart UML diagram will be model blocks and states, data blocks and processes. Clicking on data blocks will take the user to the source data used in the process. All the data used in the process would ideally reside in Romulus and the links will be URL links to relevant data sections in Romulus.
The user will be able to navigate from all model related pages to a page that elaborates the mathematical model used in the process. The model will be represented mathematically and graphically where possible. The user will be able to link back to the data used in the model from this page, as well as to the posterior distribution produced by the model. The user will be able to navigate back to the big-picture state-chart diagram with a single button-

click from any of the pages related to the model. It is hoped that this back and forth linking will help users easily understand the context of individual components.
UML Activity diagrams will also be useful for showing more detailed process flows, and Use Case diagrams may be useful for outlining the relationships between processes and the users.

f. **Structured list**: An alternative to the flow-chart is a bulleted list, where the process is outlined sequentially in order. Different levels of bullets can be used to represent sub-processes. The user will be able to click on individual items related to models or data and view the associated information as described in the previous section. This is a non-graphical way of displaying the information communicated using state-chart diagrams.

6. ARCHITECTURE

6.1 General Approach

Nesstar already provides a context in which statistical data and results can be viewed and manipulated. Associated with this can be metadata and interpretation, accessible both directly with the results and via hyperlinks. This functionality is implemented as a set of components which can be displayed in a browser interface (the Nesstar Explorer). A number of standard displays are available, which can be driven directly from the loaded data. Additional (more specialised) displays can be constructed using the components, and where a particular type of display is needed often, a template can be constructed and reused.

We will follow the same general structure for the implementation of the Provenance and Reliability functionality to be implemented in WP6.

Templates will provide the basic structure for all the different types of display of the different types of information from the model metadata. Templates will make use of components for the various types of specialised displays. Wherever possible, components will be sourced from existing functionality. For example, the R system has a comprehensive set of facilities for the display of uni- and multi-variate distributions, both formally and empirically determined. MathML can display mathematical formulae in a browser context, Visum can display traffic networks, and a SVG display component could be used for graphs.

The metadata for the fitted models will be generated (or otherwise captured) during the model fitting process. The full details of the structure are still to be elaborated, but the metadata storage will almost certainly be in XML documents. These can be manipulated using the standard set of XML-related technologies.

In the next section, an overview of specific technology components that will make up the architecture is provided along with their suitability to serve the required need. This is following by a description of the high-level architecture describing how these technology components fit together to form a working software application.

6.2 Technology components

6.2.1 Nesstar

Nesstar already provides a mechanism to store statistical data and its associated meta-data (using the DDI structure). It is envisaged that generated data and the associated meta-data will be stored in the Nesstar system and made available through web-pages powered by Nesstar in the same way. The data on Nesstar will be linked to associated information hosted outside Nesstar through URL links and vice versa.

6.2.2 R and Rserve/RCgi

R is a language and environment for statistical computing and graphics, which provides a rich graphical functionality for representing statistical data. This feature of R

will be used for graphic display of distributions and their comparison. However, the R environment needs to be made accessible to the web based system used to display OPUS related data and processes. This will be accomplished by the use of Rserve [Rserve], which will enable the R environment to act as a server that can be accessed through TCP/IP. An alternative approach to display R graphics over the web is through Rcgi [Rcgi], which provides a web interface to R. Graphics generated by R will be web-enabled by using either Rserve or Rcgi; the decision will be made after analysing the pros and cons of both technologies. Our preferred approach at this point after an initial analysis is to use Rserve, and rest of the architecture is presented assuming the use of Rserve.

6.2.3 HTML and XHTML powered by Apache

All the web-pages displaying provenance and reliability information will be coded in HTML and XHTML and hosted on an Apache web-server. The web pages will have links to Nesstar pages and there will be links from Nesstar pages to this site, thereby providing a seamless user interface through a web browser.

Provenance information will be displayed using a combination of diagrams (probably powered by SVG) and structured webpages in HTML. If R+Rserve is used to display R generated graphics, then Apache Tomcat [Tomcat] will run in conjunction with Apache web server to provide the necessary Java interface for Rserve.

6.2.4 VISUM

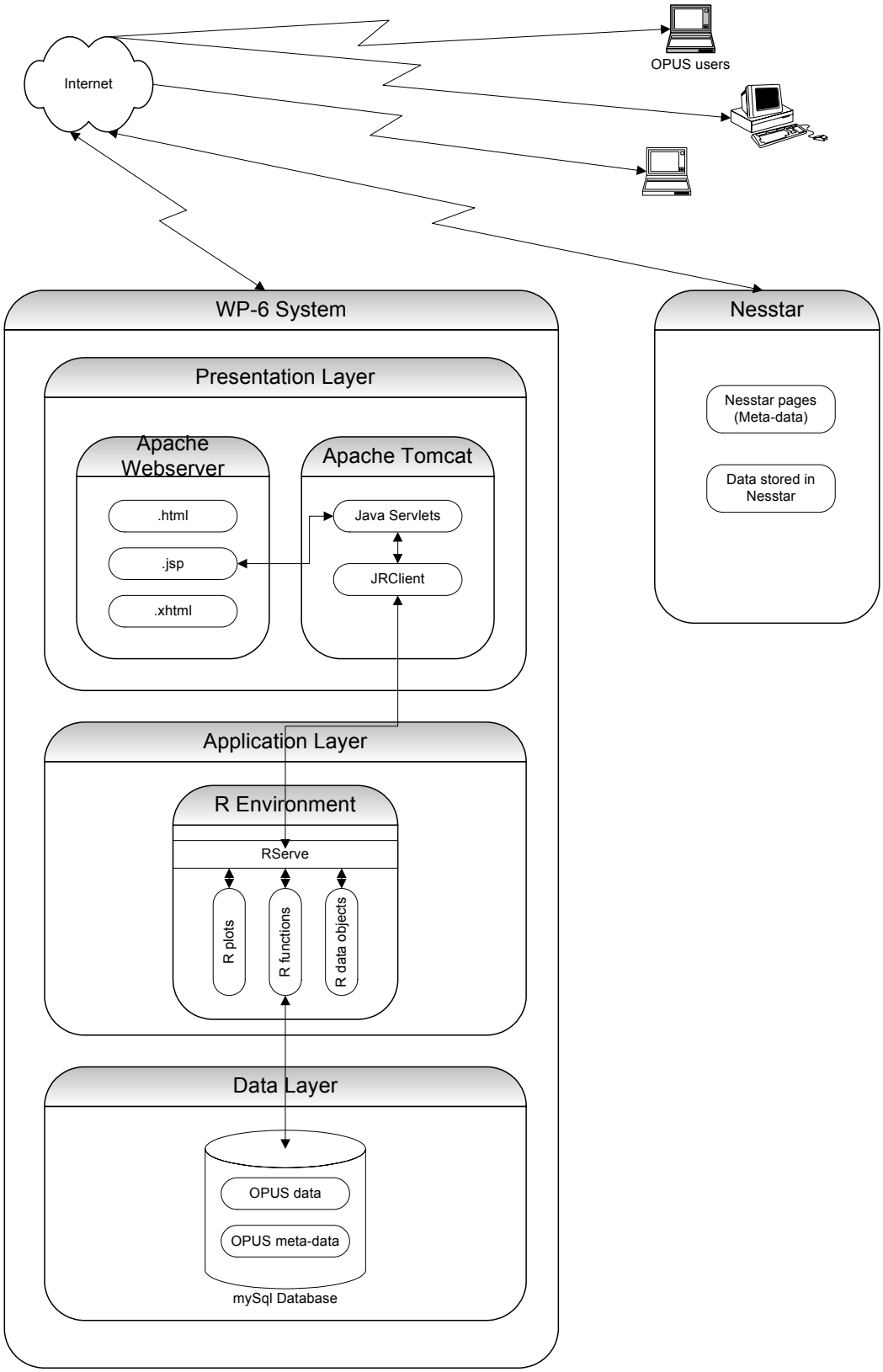
The spatial aspect of synthetic data is displayed through VISUM. However, we will rely on R for graphics representing statistical properties of data. It is proposed that the link between VISUM and R will be provided through the native C interface of R. Technical details of the integration between VISUM and R will be discussed with PtV and will be implemented as a part of WP07.

6.3 High-level architecture

This section outlines how the technology components outlined in the previous sections fit together to form a software system that will support OPUS users in viewing and understanding the models and synthetic data generated by the methodology. This section does not talk about the specific displays built as a part of this work-package, but it talks about the building blocks and the framework that will be used to build specific displays.

6.3.1 Web based system

The basic framework that combines the technology building blocks of the system is shown in the diagram below. The primary technology option for the framework involves the use of Rserve, and the logical architecture depicts this option. The secondary option involves the use of Rcgi, and this option will be used as an alternative architecture in case of any unforeseen technology issues with the primary option.



The html pages will be served by an Apache web server [Apache]. The JSP pages will be constructed dynamically where necessary. Graphics and data required for building the pages will be requested by java servlets running inside the Tomcat environment, and will be served by the R environment. The communication between the servlets and the R environment is enabled by JRClient and RServe. The R code run-

ning inside the R environment will in turn retrieve necessary data stored in a mySql database.

The above logical architecture is designed in such a way that the technical design of the system follows an MVC (Model-View-Controller) pattern. The jsp pages are responsible for the display of the information. The flow between jsp pages will be controlled by Java servlets, which also will be responsible for pulling together the required information for creating a web page from R. The R environment forms the model in the MVC architecture, where any statistical computation is carried out and the necessary plots and graphs are created.

The above architecture is designed with scalability in mind. The three logical layers can be deployed on separate physical machines, should the system need to serve a large number of concurrent users in the future. Apache web servers can be clustered, and can be load balanced across multiple machines running R to provide further scalability if required.

6.3.2 Integrated VISUM-R environment

R provides a C API which is well documented on the R site. VISUM is built using C and the C API of R can be used to integrate VISUM with R. It is envisaged that the necessary data and commands will be passed from VISUM to R. R will process the data and generate the necessary graphics, which will be displayed by VISUM. The interface design as well as the technical design of combined VISUM-R functionality will be explored and implemented as a part of Work Package 7.

REFERENCES

[Apache] Apache HTTP Server Project. Available at http://httpd.apache.org/

[DDI] Data Documentation Initiative, a standard for statistical metadata. See DDI Alliance, www.ddialliance.org.

[Fowl04] UML Distilled, 3rd Edition. ISBN: 0 321 19368 7. This is an excellent though terse guide to the content and use of UML, aimed at readers with some programming background.

[FWdV03] The Concept of Statistical Metadata (2003) by Froeschl, Grossmann, Del Vecchio, a deliverable from the MetaNet project, at www.epros.ed.ac.uk/metanet/deliverables/deliverables.html

[GCSR04] Bayesian Data Analysis. Gelman, Carlin, Stern & Rubin, Chapman Hall, 2004

[HøDe04] The gRbase Package for Graphical Modelling in R by Søren Højsgaard and Claus Dethlefsen. See www.math.aau.dk/research/reports/reports.htm, report R-2004-19.

[MathML] Mathematical MarkUp Language. An XML language for the display and evaluation of mathematical expressions. See www.w3.org/Math.

[MetaNet] See www.epros.ed.ac.uk/metanet.

[R] The R project for Statistical Computing. See www.r-project.org/index.html.

[Rcgi] Web interface for R and Octave. Available at http://www.ms.uky.edu/~statweb/

[RServe] Interactive Software developed at RoSuDa – Rserve. Available at http://stats.math.uni-augsburg.de/Rserve/

[SVG] Scalable Vector Graphics. See www.w3.org/Graphics/SVG/About.html

[Tomcat] The Apache Jakarta Project. Available at http://jakarta.apache.org/tomcat/

[UML] The Unified Modelling Language. See www.uml.org for information about UML 2.0. This is a standard developed under the auspices of the Object Management Group (www.omg.org).

[West01] LATS Database Design Study: Database Design Report, by Andrew Westlake. Available at www.sasc.co.uk/Projects/LATS DDR Public.pdf.

[West02] XML and Standards, by Andrew Westlake. Available at www.sasc.co.uk/Guides/XML%20and%20Standards.zip.

[West03] Models and Metadata, by Andrew Westlake. See www.sasc.co.uk/Guides/models%20and%20metadata.htm.

INDEX

Audit trail, 6, 17, 19, 22, 30

Bayesian, 10

Conclusions, 5, 18, 22

Confidence, 11, 20

Data
  Real, 17, 19
  Synthetic, 6, 14, 16, 17, 18, 19, 20, 22

DDI, 5, 15, 17, 21, 26, 32, 36

Display
  Comparison, 29
  Component, 24, 29

Displays
  Domain, 2, 21
  Generic, 21
  Specialised, 21, 32

Distribution
  Compare, 29

Estimates, 5, 6, 8, 9, 11, 18, 19, 20, 22

Metadata
  DDI, 17, 21, 26
  Model, 17, 21, 23, 26, 32
  Statistical, 12, 14, 36

Model
  Bayesian, 3, 28
  Checking, 28, 29
  Statistical, 5, 6, 7, 16, 17, 19, 22, 23, 24, 25, 26, 27, 28
  Transport, 20, 23, 27

mySql, 35

Precision, 11, 20, 21

Provenance, 2, 5, 6, 8, 13, 16, 19, 20, 26, 29, 30, 32, 33

R, 3, 4, 21, 23, 24, 28, 29, 32, 33, 34, 35, 36

Reality, 19, 20

Reliability, 2, 3, 5, 6, 8, 13, 16, 19, 20, 21, 25, 26, 28, 29, 30, 32, 33

Sample
  size, 19, 20

UML, 20, 27, 28, 30, 36

Uncertainty, 5, 6, 7, 9, 18, 19, 20

Variability, 6, 9, 19, 20, 21

XML, 15, 17, 23, 32, 36